Appendix B – Outlier and Influential Observation Detection

This section provides equations for determining outliers and influential observations when using regression to predict a total from a sample. In particular, assume that the operators in population are arranged in descending order of natural gas production, labeling them from $i = 1,..., N$. The first “n” operators are selected to be in the (cut-off) sample. x_i is the volume of gas produced by company i during a previous time period. x_i is assumed to be known for all operators in the population. y_i is the volume of natural gas produced by company *i* during the current time period. It is assumed that these data follow the model: $y_i = \beta x_i + e_i$. Then $\hat{\beta}$ is estimated either with ordinary least squares ($V(y_i) = \sigma^2$), or weighted least squares ($V(y_i) = \sigma^2 x_i$), providing an estimate for production in the current time period for operators not in the sample, $\hat{y}_i = \hat{\beta} x_i$ ($i = n+1,..., N$).

Hence the total gas production in the current time period is estimated as

$$T = \sum_{i=1}^{n} y_i + \hat{\beta} \sum_{i=n+1}^{N} x_i$$

The first section below describes the procedures of using ordinary least squares. The second section below describes the procedures for using a specific version of weighted least squares. It is this second section that leads to the ratio estimator we have been discussing. Hence, we recommend that RPD test the equations provided in the second section.

1. Ordinary Least Squares ($V(y_i) = \sigma^2$).

a. Detection of outliers.

The residual for company *i* is given by $e_i = y_i - \hat{\beta} x_i$.

Where $\hat{\beta} = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^{2}}$

The influence for company *i* is $p_{ii} = \dfrac{x_i^{2}}{\sum_{i=1}^{n} x_i^{2}}$.

The internally studentized residual (pg. 74) is $r_i = \dfrac{e_i}{\hat{\sigma}_1 \sqrt{1 - p_{ii}}}$. The studentized residuals follow a standard normal distribution (if the individual data follow a normal distribution).

This fact is used to determine suitable cut-off values for classifying an observation as an outlier. We have been using 3.5 to define outliers. If the data actually follow a normal distribution this will only classify about 0.05 percent of the observations as outliers. However, in our experience, many more observations than this are actually classified outliers using the EIA-23 data. The data reported by company *i* is an outlier if

$$|r_i| > 3.5$$

b. Detection of influential observations.

The externally studentized residual is defined to be $r_i^* = \dfrac{r_i\sqrt{n-2}}{\sqrt{n-1-r_i^2}}$. This version of the studentized residual can also be used to determine outliers. However, it is most important as an intermediate value in computing DFFITS, a measure of influence. An observation is influential if it has a large impact on the estimated value of β.

$$DFFITS = |r_i^*|\sqrt{\frac{p_{ii}}{1-p_{ii}}}$$

A rule of thumb for classifying observations as influential is $DFFITS > \dfrac{2}{\sqrt{n}}$.

Observations that are classified as outliers or influential are excluded from the sample for purposes of computing β. Their data are still used in estimating the total production (after verification by survey staff that the data are accurate.)

2. Weighted least squares $(V(y_i) = \sigma^2 x_i)$

With the equations below the estimated total can be rewritten as follows:

$$T = \sum_{i=1}^{n} y_i + \hat{\beta}\sum_{i=n+1}^{N} x_i = n\bar{y} + \frac{n\bar{y}}{n\bar{x}}(X - n\bar{x}) = \frac{n\bar{y}}{n\bar{x}}X = \hat{\beta}X$$

or

$$T = \frac{X}{n\bar{x}}n\bar{y}\,.$$

a. Detection of outliers.

For outlier and influential detection redefine the following terms.

$$e_i = \frac{y_i - \hat{\beta}x_i}{\sqrt{x_i}}\ ,$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} \ ,$$

$$p_{ii} = \frac{x_i}{\sum_{i=1}^{n} x_i} \ ,$$

$$r_i = \frac{e_i}{\hat{\sigma}_2 \sqrt{1 - p_{ii}}} \ ,$$

and $\sigma_2{}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(y_i - \hat{\beta} x_i)^2}{x_i}$.

As noted above, we have been classifying the data reported by company *i* as an outlier if $|r_i| > 3.5$

For computing the influence statistic, compute the same equations as for ordinary least squares but use the new definitions of r_i and p_{ii} from this section.

$r_i^* = \frac{r_i \sqrt{n-2}}{\sqrt{n-1-r_i^2}}$ and $DFFITS = |r_i^*| \sqrt{\frac{p_{ii}}{1 - p_{ii}}}$.

b. Detection of influential observations

A rule of thumb for classifying observations as influential is $DFFITS > \frac{2}{\sqrt{n}}$.

Observations that are classified as outliers or influential are excluded from the sample for purposes of computing β . Their data are still used in estimating the total production (after verification by survey staff that the data are accurate.) Hence the estimate for the total would be written as follows:

$$T = \sum_{i=1}^{n} y_i + \hat{\beta} \sum_{i=n+1}^{N} x_i = n\bar{y} + \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} (X - n\bar{x})$$

Note that with outliers and influential observations removed, the equation does not simplify quite as nicely. $i \in s$ denotes sampled data that are not outliers or influential observations. $n\bar{x}$ and $n\bar{y}$ denote the total from the sampled operators in the previous time period and the present time period respectively – these totals include influential observations and outliers.